Received: December 3, 2024. Revised: January 23, 2025. 650

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

FER-MOTION: Facial Expression Recognizer in Multi-resolution Images Using a Lightweight Large Receptive Residual Network

Muhamad Dwisnanto Putro[1*] Wahyono[2] Joko Hariyono[3]
Oktavian A. Lantang[1] Danilo Cáceres Hernández[4]

[1]*Department of Electrical Engineering, Faculty of Engineering, Sam Ratulangi University, Manado, Indonesia*
[2]*Department of Computer Science, Gadjah Mada University, Yogyakarta, Indonesia*
[3]*Department of Electrical Engineering, Sebelas Maret University, Surakarta, Indonesia*
[4]*Facultad de Ingeniería Eléctrica, Universidad Tecnológica de Panamá, Panamá*
*Corresponding author's Email: dwisnantoputro@unsrat.ac.id

Abstract: Real-world applications challenge facial expression recognition systems to adapt to various input image resolutions. Specifically, two-stage methods that rely on face patches from face detection tend to produce limited information for low-resolution cases and slow in the inference stage for high-resolution input. Besides, a high-performance facial emotion vision-based system requires an adaptive deep learning model with low parameter usage and computational cost. This work proposes a novel facial emotion recognizer in multi-resolution input (FER-MOTION) with high performance and cost-efficiency. The proposed network offers a lightweight CNN approach that is improved from MobileNetV2, offering a large kernel receptive module and a pyramid enhancement module, each designed to improve effectiveness and efficiency. This approach introduces a new extractor module capable of discriminating facial emotion features in a lightweight operation by capturing a larger spatial area at each network stage. A group-based attention module involving a pyramid spatial map is proposed to overcome the saturation performance of the extraction network. Comprehensive experimental results demonstrate that the proposed CNN architecture achieves high accuracy across varying image resolutions. The experiment is conducted on three benchmark facial expression datasets: KDEF, RAF-DB, and FERPlus. Analyses and comparisons of computational and parameter efficiency show that the proposed model is 3.8 times lighter in parameters and 1.8 times more efficient in floating-point operations than MobileNetV2.

Keywords: FER-motion, Facial expression, Lightweight model, Multi-resolution, Enhancement module, Low-cost computation.

1. Introduction

The challenge in multi-resolution facial expression recognition occurs when the model processes low-resolution images. Key features like the eyes and lips become too small to be accurately identified, making it difficult for the model to recognize expressions effectively [1]. Facial expression recognition is also challenging due to the significant variability in human faces, which complicates the model's ability to detect expressions consistently. Furthermore, the subtle differences between similar expressions, such as disgust and anger or shock and surprise, make it difficult for the model to distinguish the expression accurately [2]. Computational efficiency issues can be achieved by processing data at lower resolutions and switching to higher resolutions only when necessary to save computing time and resources [3-5]. Clustering low-resolution information involves recognizing patterns in critical features, enabling the system to group data more effectively while preserving important details for later analysis. This approach enables better adaptation to rapidly changing conditions. It also allows the system to customize more adaptive responses without processing the entire data at full

resolution. In addition, many data types naturally exhibit a hierarchical structure, with basic features present at lower resolutions and more complex features emerging at higher resolutions [6]. Employing a multi-resolution approach can leverage this hierarchy for more effective analysis and decision-making. In computing, multi-resolution techniques also optimize data representation in real-time processing by integrating information across different resolution levels [7, 8]. Therefore, the proposed system will adjust the requirements of the deep learning model based on the dimensions of the detected face patches.

A Convolutional Neural Network (CNN) is a deep learning method to process structured image data [9-13]. In applications such as human-computer interaction and the Internet of Things (IoT), it is crucial to utilize CNN architectures with a reduced number of parameters [14, 15]. It is particularly well-suited for devices with limited capabilities, ensuring accurate predictions even with multi-resolution images [16, 17]. CNN utilizes multiple channel layers to discriminate essential features. Shallow layer architectures typically apply fewer convolution operations than deep architectures such as Visual Geometry Group (VGG) [18] and Residual Networks (ResNet) [19]. The lack of a superficial network captures complex and challenging features [20]. An enhancement module offers a solution to address this issue by implementing a comprehensive attention block that highlights the essential features of the input feature map [21]. This process designs trained weights to refine input features and improve the prediction system in machine learning [22]. The proposed work introduces two enhancement modules, including a spatial enhancement module assigned to critical parts of the large receptive residual and a pyramid-based enhancement module that processes the final features of the backbone and recovers valuable features for decision-making in the classifier block. The assignment of the two modules avoids excessive computation and parameters, thereby increasing the efficiency factor of the model. The contributions of this work are summarized as follows:

1. A novel facial emotion recognition (FER-MOTION) that utilizes lightweight and computationally efficient CNN architecture is introduced for recognizing basic facial expressions in multi-resolution input images. The Large Receptive Residual Network (LR2) captures a larger spatial area of the input map, enhancing the variety of combined representations. This architecture employs efficient operations to maintain low computational costs and high performance.
2. Two novel enhancement modules: A spatial Context Enhancement module that strengthens the relationship of channel information within a single spatial map and an Efficient Pyramidal Enhancement module (EPE) highlight vital features across different groups and spatial areas.
3. Extensive experiments are conducted on multi-resolution input images involving three benchmark datasets for facial expression recognition: Karolinska Directed Emotional Faces (KDEF), Real-world Affective Faces Database (RAF-DB), and Facial Expression Recognition 2013 Plus (FERPlus). The study also evaluates the model's efficiency, comparing parameter usage, computational complexity, and speed on Jetson Nano and low-computation Central Processing Unit (CPU) devices. In addition, real application scenarios visually test the reliability of the integrated system when implemented on an inexpensive device.

The structure of this paper is as follows: Section II provides an overview of the related studies in facial expression recognition. Section III outlines the methodology, detailing the FER-MOTION system's architecture and key modules. Section IV describes the experimental setup, including datasets and evaluation criteria. Section V discusses the results and their implications. Lastly, Section VI concludes the paper and suggests directions for future research.

2. Related works

Facial expression recognition in real-world scenarios encounters significant challenges, such as occlusion, varying resolutions, and pose variations, which can impact accuracy. Advanced methods have been developed to address these issues, such as sparse autoencoders for facial expression recognition [23]. It was implemented for pain assessment applications by focusing on the upper part of the facial image to mitigate the challenges posed by different poses. Another study implemented ResNet-50, modifying the residual down-sampling block to enhance facial expression recognition [24]. This method demonstrated high accuracy across several benchmark datasets. More recently, Graph Convolutional Networks (GCNs) have been developed, introducing High Aggregation Subgraphs (HASs) [25]. Modern deep learning methods incorporate transformer models that leverage self-attention to establish global relationships between features, as demonstrated by [26-27]. In addition, a combination of approaches is particularly adept at handling image recognition at large input resolutions, consistently achieving satisfactory accuracy in their

respective challenges. For instance, a study [28] implements an ensemble feature extractor with MDNet, ViT, and pre-trained ResNet50 to improve accuracy and precision in recognizing mental disorders.

Low-resolution input images can hinder performance due to limited information availability [29]. Recent studies have addressed this challenge by employing lightweight convolutional networks that maintain high effectiveness [30]. Probabilistic data uncertainty learning has also been applied to enhance feature learning by focusing on erroneous predictions caused by input feature constraints [31]. The challenge of low-resolution inputs lies in the reduced feature variation, making it difficult to apply these methods to samples in the wild. A residual voting network, which modifies ResNet-18, has been proposed to address this issue. It can improve the focus on critical features and reduce resource consumption by minimizing the extraction area.

Many previous studies have addressed facial expression recognition with varying focuses and limitations. The study [23] concentrated solely on the challenge of pose variation, which limits its adaptability to other significant factors like resolution changes. Studies [29-31] focused exclusively on low-resolution inputs, achieving good performance in such cases but lacking generalization to higher resolutions. Conversely, a study [32] targeted high-resolution images, resulting in inefficiencies when applied to lower-resolution scenarios. These approaches collectively highlight the rigidity of existing models, which are constrained to specific resolution types and fail to adapt to significant dimensional differences. Furthermore, studies such as [24, 25, 32] overlooked practical application aspects, particularly efficiency. In contrast, the FER-MOTION architecture is specifically designed to overcome the limitations of previous work [23], which struggles with resolution changes. FER-MOTION is highly adaptable and performs effectively across high and low-resolution inputs, ensuring consistent and flexible performance in multi-resolution scenarios. Compared to studies [26-32] that only work well on specific resolution images, FER-MOTION bridges this gap by supporting a wide range of resolutions to reach implementations on objects of varying scales that often occur in real case scenarios. Additionally, while studies like [24], [25], and [31] overlook efficiency, FER-MOTION is lightweight, using fewer parameters and less computational power. Despite its simplicity, it delivers high accuracy while remaining adaptable and efficient for real-world applications involving diverse resolutions.

3. Method

The proposed deep learning system uses a data-driven feature extractor to distinguish unique facial information. It emphasizes the importance of correlating vital features the classifier can use to generate accurate facial expression predictions on particular assessments. The FER-MOTION network is designed to maximize the smoothness of variation by dividing the process into two main stages, each handled by a dedicated module. The architecture consists of two core modules, the backbone and the classifier, as shown in Fig. 1.

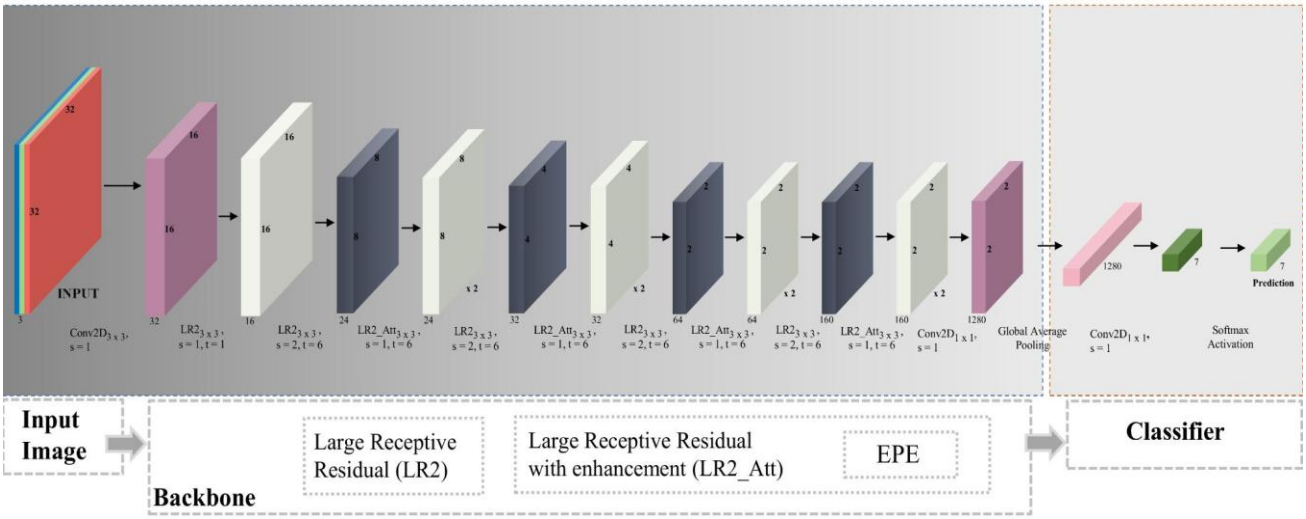


Figure. 1 Overall architecture of FER-MOTION. The backbone integrates Large Receptive Residual and Large Receptive Residual with Enhancement modules interchangeably to effectively distinguish essential facial features from irrelevant information. The EPE (Efficient Pyramid Enhancement) module is embedded within the LR2 with Enhancement to further augment the feature extraction capability

3.1 Backbone

The backbone module is inspired by MobileNetV2 architecture [33], where the backbone module aims to extract the main features of expression. It utilizes convolution operations that are optimized for computational efficiency. It allows for more precise identification of important facial features. Once these features are extracted, the classifier module assumes the task of generating the final prediction of the facial emotion expressed based on the features extracted by the backbone. The FER-MOTION architecture can improve accuracy in facial emotion recognition.

The backbone consists of a convolution layer, two proposed modules, Global Average Pooling, and softmax activation. The proposed two modules include Large Receptive Residual (LR2) and Large Receptive Residual with attention mechanism (LR2_Att). Large Receptive Residual is utilized as a bottleneck replacement in MobileNetV2 and aims to extract features using a 5 × 5 kernel depthwise convolution. Employing a large kernel provides feature extraction with a sizeable receptive region and efficient computation due to the utilization of depthwise convolution. Additionally, LR2 employs a stride of two to reduce the spatial dimensions.

Furthermore, 1 × 1 convolution with SmoothSwish activation [30] was applied. This study employs this simple activation function for each LR2 operation, preserving small features. This approach prevents and controls the loss of regions under the negative curve, which can occur due to convolution operations and the use of bias. The information represented in these regions typically contains subtle facial gesture features, and their removal can lead to decision errors. Therefore, the study applies a beta value of 1.2 to maintain the score space in small negative regions while forcing large negatives to disappear, thereby minimizing their impact on neighboring features. Additionally, a 1 × 1 convolution with batch normalization is applied and residualized with the initial input to restore the features lost due to the convolution process (x). The mathematical representation for LR2 is as follows:

$$LR2 = (BN(Conv_{1x1}(\delta(Conv_{1x1}(BN(Dwl_{5x5}(x))))) + x \quad (1)$$

where x is the input feature that belongs to Dwl, and Dwl is a depthwise convolution with a large kernel size of 5 × 5, ensuring a sizeable receptive area. BN represents batch normalization, which normalizes the extracted weight values. $Conv_{1x1}$ represents an ordinary convolution with a 1 × 1 kernel size, and δ indicates SmoothSwish activation utilized to transform integer values into probabilities.

3.2 Large receptive residual with enhancement module

The backbone is designed as a crucial block that distinguishes essential expression features from trivial information. The FER-MOTION network leverages standard convolution's extensive information retrieval capabilities to efficiently capture a larger feature area. However, low-computation operations often miss high-level features of sufficient quality. This problem decreases network performance. In order to address this case, this study incorporates a spatial-based attention module within the down-sampling convolution block. This addition aims to enhance the beneficial features while mitigating excessive feature loss that typically occurs during sequential convolutional blocks [21]. Fig. 2 illustrates the integration of an enhancement module within the residual part to increase feature correction performance. This attention mechanism effectively captures valuable features across the entire spatial area of the input map without significantly increasing the number of parameters. The formulation of the LR2 with attention module is as follows:

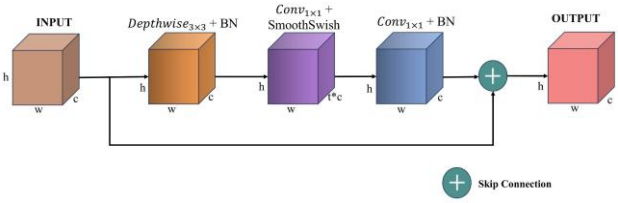


Figure. 2 The proposed Large Receptive Residual module. This module employs sequential lightweight operations and integrates a residual technique to identify and recover missing features at the final process of the operation

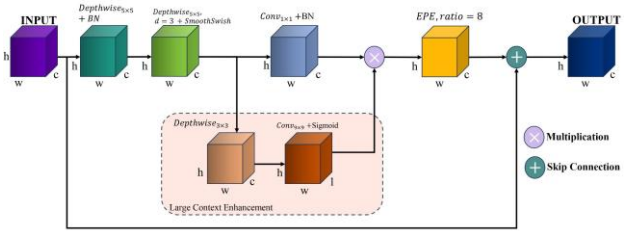


Figure. 3 The proposed Large Receptive Residual with Enhancement Module. This module incorporates an attention mechanism following two large depthwise convolutions to emphasize valuable information

$$LR2_{att}(x) = EPE\left(BN\left(\begin{matrix}Conv_{1x1}(Ds)) \otimes \\ (LCE(Ds))\end{matrix}\right)\right) + x \quad (2)$$

where

$$LCE = \sigma(Conv_{9x9}(Dw_{3x3}(Ds))) \quad (3)$$

and

$$Ds = \delta(Dw_{5x5}(Dw_{5x5}(x))) \quad (4)$$

The proposed block applies two attention modules: spatial attention (*Ds*) and efficient pyramid enhancement (*EPE*). The attention-based LR2 module applies a combination of depthwise operations with a 5 × 5 kernel (Dw_{5x5}), applying dilation of three on the last convolutional. This operation was followed by smooth swish activation (δ). Furthermore, a Large Context Enhancement (*LCE*) module is offered to improve the representation of vital features in spatial regions by applying a depthwise operation with a 3 × 3 kernel followed by a more extensive filter operation ($Conv_{9x9}$) to generate a single map. A sigmoid activation (σ) generates a weighted probability map to refine the constructed features from $Conv_{1x1}$. This enhancement can strengthen the backbone when passing the extracted information to the network. Although it uses a large filter, it is a lightweight module that only generates a single weighted map.

3.3 Efficient pyramid enhancement

The medium layer of the backbone provides complex features produced from a comprehensive sequence of convolution operations. This feature map contains coarse facial gesture information; additional mechanisms are required to refine the valuable features. Several small and significant intensities of facial elements need to be enhanced to force the classifier's ability to strengthen. Therefore, an additional attention module is offered to achieve a high-performance backbone. On the other hand, static spatial-based operations generate saturation performance that does not relate essential features far apart [22]. The FER-MOTION proposes a spatial enhancement pyramid that can capture essential features with various receptive variations and summarize them as a valuable representation for information updating, as shown in Fig. 3. The proposed operation ignores the massive computation and thus does not significantly slow down the feature extraction calculation. An Efficient Pyramid Enhancement (EPE) applies depth-wise operations with various dilations involving the pyramid receptive regions, which is illustrated in detail as follows:

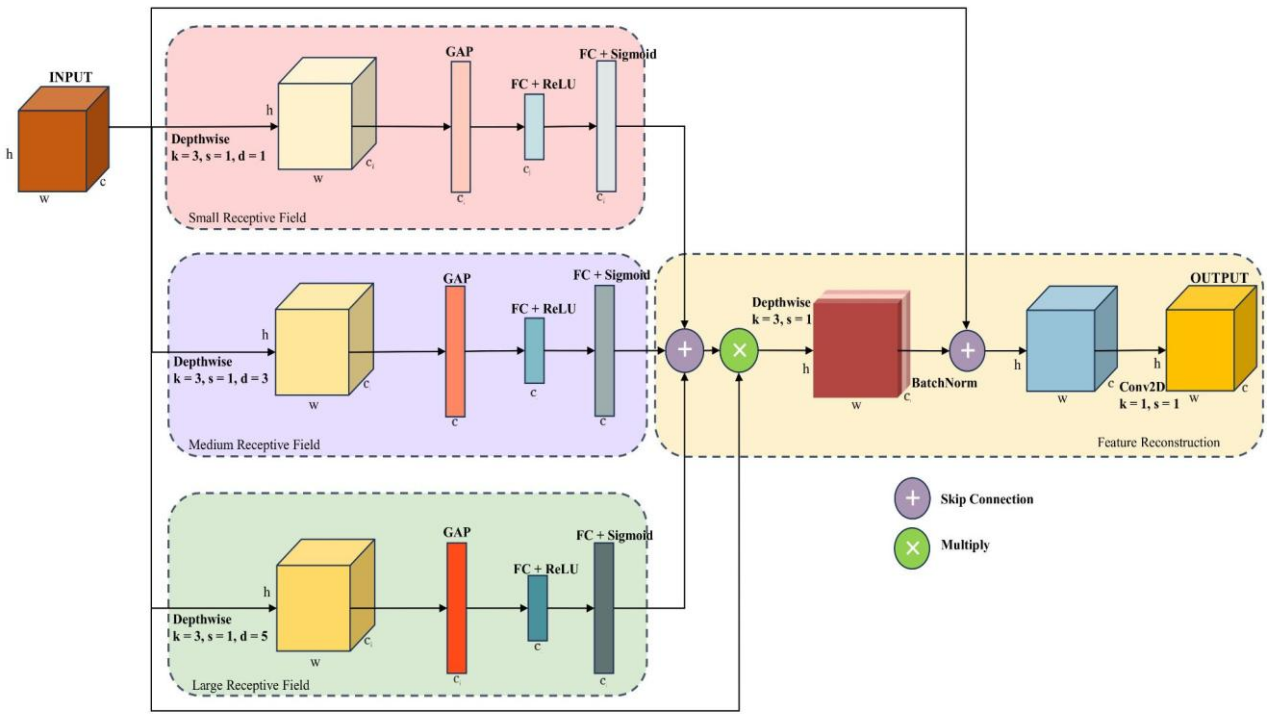


Figure. 4 The proposed Efficient Pyramid Enhancement

$$EPE = Conv_{1x1}(Dw_{3x3}(\sum_{i=1}^{n=3} Rf_i(x) \otimes x) + x) \quad (5)$$

where Dw_{3x3} is a depthwise convolutional layer to reconstruct enhanced features from the combined spatially varying features (Rf_i) . $Conv_{1x1}$ is a convolutional layer with a 1 × 1 filter to mix the information from the residual feature correction function. The feature enhancement variation utilizes three different layers of spatial representation (Rf_1, Rf_2, and Rf_3) formulated as follows:

$$Rf_1 = \sigma(Z_{12}(\gamma(Z_{11}(GAP(Dw_{3x3,d=1}(x)))))) \quad (6)$$

$$Rf_2 = \sigma(Z_{22}(\gamma(Z_{21}(GAP(Dw_{3x3,d=3}(x)))))) \quad (7)$$

and

$$Rf_3 = \sigma(Z_{32}(\gamma(Z_{31}(GAP(Dw_{3x3,d=5}(x)))))) \quad (8)$$

The combination of the receptive region involves a channel-based attention operation that adopts work from [21]. Each extracted receptive difference map is summarized by a Global Average Pooling (*GAP*) operation to generate the vector-based excitation scores. Then, two fully connected layers are sequentially employed, Z_{i1} and Z_{i2}, to select the preferred channel information followed by ReLU (γ) activation. The final block applies sigmoid activation (σ) to produce a weighted vector to refine the input map. The variable notation for all equations is presented in Table 1. The weights generated by each attention are summed by an element-wise addition operation combining the features at each position of the same vector.

3.4 Classifier and loss function

In the proposed network, a 2D convolution with a 1 × 1 filter generates 1,280 features, representing the number of channels. A global average pooling summarizes the features map by taking the mean of each score channel, thus preventing parameter overload. In addition, the 2D-convolution operation creates vectors with dimensions corresponding to the number of predicted emotion categories. This task uses seven emotion classes on the KDEF and RAF-DB datasets and generates eight categories on the FERPlus dataset, representing basic human facial expressions. The last layer applies softmax activation to produce probabilities associated with a multimodal distribution, allowing the model to control a

Table 1. Notation List

Variables	Description
x	Input
Dwl_{5x5}	Depthwise convolution with a large kernel size of 5 × 5
BN	Batch normalization
$Conv_{1x1}$	Ordinary convolution with a 1 x 1 kernel size
δ	SmoothSwish activation
$LR2$	Large Receptive Residual
Ds	Spatial attention
LCE	Large Context Enhancement
EPE	Efficient Pyramid Enhancement
$LR2_{att}$	Large Receptive Residual attention
Dw_{3x3}	Depthwise Convolution with a 3 x 3 kernel size
$Conv_{9x9}$	Convolution with a 9 x 9 kernel size
σ	Sigmoid activation
Dw_{5x5}	Depthwise Convolution with a 5 x 5 kernel size
GAP	Global Average Pooling
$Dw_{3x3,d=1}$	Depthwise Convolution with a 3 x 3 kernel size and 1 dilation
$Dw_{3x3,d=3}$	Depthwise Convolution with a 3 x 3 kernel size and 3 dilation
$Dw_{3x3,d=5}$	Depthwise Convolution with a 3 x 3 kernel size and 5 dilation
Z_{i1}	Fully Connected
γ	ReLU activation
Z_{i2}	Fully Connected
Rf_1	First layers of spatial representation
Rf_2	Second layers of spatial representation
Rf_3	Third layers of spatial representation

Table 2. Ablation studies of the proposed module evaluated on KDEF dataset

Module	Experiment				
	1	2	3	4	5
S-swish	✓				
EPE	✓	✓			
LCE	✓	✓	✓		
Additional Depthwise	✓	✓	✓	✓	
Eff MobileNetV2	✓	✓	✓	✓	✓
Parameters	595,549	595,549	491,479	440,511	1,108,863
Accuracy (%)	97.96	97.46	96.28	95.82	94.78

wide range of facial expression predictions. Furthermore, the proposed network employs a sequential convolutional block that uses a categorical cross-entropy loss function to compute the prediction loss of multi-resolution face expression recognition. The method compares the model's output with predefined ground truth labels involving one hot label. It can allow the evaluation of the model's performance in recognizing faces at different resolutions.

3.5 Implementation setup and datasets

The training stage increases model ability through a learning feature process that involves Adam Optimizer with a momentum of 0,9. Hyperparameters for fine-tuning include cosine learning rate decay, a batch size of 128, and no weight decay. A learning rate of 0.01 is used, and training is performed for 100 epochs. During the inference phase, the system crops the detected face region and feeds it into the classification model. This real-application experiment used a Logitech c270 webcam with VGA resolution (640 × 480) for live stream input. The classification model handles 32 × 32-pixel patches, with speed measured across 1000 frames to capture maximum performance. Experiments were carried out on both Jetson Nano to reflect affordable machine processing. The proposed facial emotion identification system involves face detection to isolate the face region from the background. This step is crucial for directing the focus of the classification model exclusively on the face, enhancing the system's performance. The proposed system employs face detection [4], a rapid and accurate method for detecting small faces in different resolutions. The proposed model integrates classification with face detection, using knowledge from the KDEF dataset in real scenario cases.

In order to train and evaluate the facial expression recognition network, several multi-resolution datasets were utilized, including the Real-world Affective Faces Database (RAF-DB) [34], Karolinska Directed Emotional Faces (KDEF) [35], and Facial Expression Recognition 2013 Plus (FERPlus) [36]. Augmentation techniques, such as rotation, flipping, brightness adjustment, contrast enhancement, and color distortion, were applied exclusively to the KDEF dataset. Linear interpolation was used to upscale and downscale images to create multi-resolution inputs. The datasets served as the knowledge base from which the proposed model learned, with no prior training applied. The trained model assessed prediction errors using categorical cross-entropy loss, comparing predicted results with actual labels. For model evaluation, the KDEF dataset was divided using 10-fold cross-validation with a batch size of 128, and each fold was trained over 100 epochs. The consistency of the model's performance was further tested on the FERPlus and RFDB datasets, following the data split configuration outlined in [30]'s research. These datasets were used to train the model over 500 epochs with a batch size 32. The Adaptive Moment Estimation (Adam) optimizer, initialized with a learning rate of 10−4, was employed. The learning rate was reduced by a factor of 0.75 whenever training accuracy plateaued for 20 epochs.

4. Experiments and results

This section presents the experimental results and evaluation of the proposed network, including an ablation study, comparisons with previous works, and an efficiency analysis. Performance and efficiency are measured using key metrics such as accuracy, parameter count, computational complexity, and model speed on low-cost devices.

4.1 Model analysis

In order to compare the performance of the proposed model through various experiments, this ablation study offers a thorough investigation into the usage of the modules proposed. These experiments evaluate the model's performance and parameter

efficiency by modifying the network's block structure and assessing each change's impact. Each step in the modification process is conducted carefully to observe whether the addition or replacement of a particular module affects the final results. Table 2 presents the results from five experiments, each testing a different combination of modules. The experiments are conducted on the KDEF dataset, which is preferred due to its balanced number of instances.

Table 1 illustrates the results of evaluating the performance and parameter efficiency of the proposed model. The first experiment is the proposed architecture, employing S-Swish Activation, EPE, LCE, Additional depthwise, and Efficient MobileNetV2. The complete model produced 595,549 parameters and achieved 97.96% accuracy. Secondly, it removes S-Swish Activation without impacting the number of parameters. However, its accuracy decreased slightly to 97.46%. Furthermore, it excludes both S-Swish Activation and EPE, produced 491.47 parameters, and achieved an accuracy of 96.28%. This experiment left only one enhancement module, LCE. Additional depthwise and Efficient MobileNetV2 can generate parameters to 440M, with a corresponding decrease in performance of 95.82%. The last experiment only employed Efficient MobileNetV2, which involved 1,108,863 parameters. However, the accuracy dropped by 1.04%. These observations demonstrate that increasing parameters can enhance the model's capacity, but it does not always lead to proportional improvements in performance. It can degrade performance when the number of parameters is excessive.

4.2 Model Evaluation on Datasets

4.2.1. Evaluation on KDEF dataset

This dataset provides 4,900 RGB (Red, Green, Blue) images based on laboratory conditions and situations containing seven basic facial emotions: fear, anger, neutral, sadness, disgust, surprise, and happiness. The original dataset includes 70 people, each posing in five different poses: straight pose, full right, full left, half left, and half right. It accommodates both male and female genders to increase the variety of human faces. The multi-resolution evaluation of the proposed model was performed on various image sizes, including 10 × 10, 32 × 32, 64 × 64, 72 × 72, 150 × 150, and 224 × 224 pixels. Table 3 shows that the FER-MOTION model achieves an accuracy of 97.81% at the resolution of 224 × 224 pixels. This performance surpasses other architectures proposed by Cho et al.[38], which achieved an accuracy of 87.09%. This study employs a local attention module that utilizes multiple convolutions to generate a position map. The module is applied at the final stage of feature extraction, aiming to enhance the quality of the extracted information.

Table 3. Performance comparison of the proposed model with previous works at different input resolutions evaluated on the KDEF dataset.

Model	Accuracy on Resolution					
	224 × 224	150 × 150	72 × 72	64 × 64	32 × 32	10 × 10
MobileNetV2	-	-	-	96.73	96.59	-
MobileNetV1	-	-	-	96.49	96.15	-
MobileNetV3 Small	-	-	-	95.33	95.58	84.64
MobileNetV3 Large	-	-	-	96.38	96.75	87.32
ShuffleNetV1	-	-	-	93.63	90.08	-
ShuffleNetV2	-	-	-	95.27	96.14	-
ResNet18	-	-	-	93.06	97.12	91.97
GhostNet	-	-	-	88.62	96.79	90.11
RGAA	93.47	-	-	-	-	-
Aly et al. [37]	-	-	96.27	-	-	-
Cho et al. [38]	87.09	-	-	-	-	-
SAFEPA [23]	-	94.29	-	-	-	-
FER-MOTION	**97.81**	**95.88**	**97.54**	**97.49**	**97.96**	**92.1**

Table 4. Performance comparison of the proposed model with previous works at different input resolutions evaluated on the RAF-DB dataset.

Model	Accuracy on Resolution					
	224 × 224	20 × 20	15 × 15	10 × 10	8 × 8	5 × 5
SCN	-	69.26	56.13	55.55	44.92	41.82
DMUE	-	73.63	70.60	62.13	55.93	46.54
RUL	-	80.63	75.65	69.17	64.06	56.16
MULR	-	81.10	77.44	70.96	66.30	59.18
Cho et al.	87.09	-	-	-	-	-
Sun et al. [39]	89.50	-	-	-	-	-
Jiang et al. [40]	88.72	-	-	-	-	-
CLCM [41]	84.00	-	-	-	-	-
FER-MOTION	**90.32**	**83.90**	**81.03**	**73.14**	**72.00**	**62.32**

However, this positioning strategy cannot guarantee comprehensive improvement of the image features at each stage, thereby reducing the reliability of the prediction results. The proposed model also outperformed the SAFEPA model [23] by 1.59% at the resolution scale of 150 × 150 pixels. This approach utilizes Sparse Autoencoders to reconstruct the upper part of the face, focusing primarily on the eyes, cheeks, and the upper portion of the nose. However, this design reduces the variability of information, limiting the model's ability to predict facial expressions accurately. Additionally, it relies heavily on decoders to produce high-quality image reconstructions, leading to over-processing and the potential loss of small yet critical expression details.

The excellence of FER-mOTION is reinforced at medium resolution, performing 97.54% at a 72 × 72 size. It outperforms Aly et al. [37], which differs by 1.27%. This competitor utilizes online learning states with CBAM to enhance the performance of bottleneck feature extraction. The enhancement module is integrated prior to the residual block, which is reported to improve the capability of sequential convolutional feature extraction. While the enhancement module contributes to improved accuracy, it is noted that the network remains bottlenecked during the image-gathering stage in the online learning environment, limiting its overall performance and hindering the achievement of higher accuracy.

Furthermore, the proposed model achieved an accuracy of 97.49% at the 64 × 64 pixels. It shows that the proposed model is superior to lightweight models, such as other MobileNet families, which only maintain an accuracy of approximately 96%. The proposed model outperforms 2.22%, 4.43%, and 8.87% higher than shuffleNetV2, ResNet18, and GhostNet, respectively. This superiority continues at lower resolutions, such as 32 × 32 and 10 × 10, where other efficient models have performance below our model.

4.2.2. Evaluation on RAF-DB dataset

This dataset provides a significant challenge in facial expression recognition that captures the facial image in the wild. It contains a variety of poses, occlusion styles, and illumination with static resolution. In order to reach a variant of image scale, it manipulates the size by implementing a bilinear interpolation approach. Table 4 presents the evaluation results of the proposed model across various input resolutions using the RAF-DB dataset. This experiment was conducted in different resolutions, 224 × 224 to 5 × 5 pixels, providing insights into the model's performance in identifying patterns at different levels of resolution. The FER-MOTION achieved the highest accuracy of 90% at the 224 × 224 resolution. Among the other models evaluated at this resolution, they obtain lower accuracy than our model. This evaluation demonstrates that the proposed model surpasses several benchmarks at the highest resolution.

The results indicate that the recent methods by Sun et al. [39] and Jiang et al. [40] underperform compared to the proposed approach. Although their models utilize complex deep learning algorithms with diverse feature extraction techniques, they struggle in occupied occlusion scenarios and varying illumination. Their enhancement modules are applied solely at the high-level feature stage, ignoring improvements at the medium and low levels, undermining overall correction capabilities. Furthermore, the approach by Jiang et al. [40] primarily focuses on efficiency, and its application in human-machine interaction focuses on efficiency, limiting its robustness to handle extreme conditions. Furthermore, The CLCM [41] achieved a performance of 84%, which is lower than the

proposed model. This approach relies on weak feature extraction, prioritizing efficiency while lacking an attention module to enhance the quality of the shallow network. Consequently, the modified MobilenetV2 model struggles to address the complex challenges posed by the RAF-DB dataset, highlighting the advantages of the proposed method in handling such scenarios.

The proposed model also maintained strong performance with an accuracy of 83% when we resolved to 20 × 20 pixels. This model continued to show a significant margin over existing models, highlighting its robustness even at lower resolutions. Furthermore, this evaluation investigated the model at the 15 × 15 resolution and achieved a performance of 81%. This result indicates that the FER-Motion is still superior to other networks. The success of our model is also demonstrated at low resolutions (10 × 10, 8 × 8, and 5 × 5), which results in an accuracy of 73.14%, 72%, and 62.32%, respectively. These results highlight the proposed model's robust performance across all tested resolutions, particularly in maintaining superior accuracy compared to other models. The comparison reveals that the proposed model outperforms other models across most resolution levels, with the highest performance observed at the high resolution. It also demonstrates substantial effectiveness at lower resolutions, making it suitable for real-world applications with varying image quality.

4.2.3. Evaluation on FERPlus dataset

This dataset is a refinement of the FER-2013 dataset, which improves the fit between images and labels. It contains eight basic emotions with several challenges widely used for evaluating facial expression recognition models. The comparison of the proposed model with several previous models across various input resolutions is presented in Table 5. This work evaluates the model in resolutions of 64 × 64, 128 × 128, 256 × 256, 5 × 5, 14 × 14, 18 × 18, and 32 × 32 pixels. The proposed model's accuracy increases to 89.46% at the 224 × 224 resolution. Jiang et al. [40] achieved a slightly higher performance than our model, with only a 0.18% difference. Although this difference is negligible, their model effectively performs on the FERPlus dataset by leveraging an RNN (Recurrent Neural Network) to capture relational patterns between facial features. Additionally, they employed extended training methods, such as transfer self-training, to incorporate prior knowledge of facial structure at the initial training stage, enhancing their model's capability. LenslessFET, and Ma et al. perform lower accuracy than our work. Ma et al. [42] employed a Multi-Layer Transformer Encoder to enhance the quality of essential features. This block is applied after combining extracted RGB and LBP features, leading to the transformer's module missing focus to capture vital gesture faces. As a result, the model struggles with specific multi-pose scenarios, achieving an accuracy of 88.81%. In contrast, LenslessFET [43] incorporates a Spectral Attention (SA) module to enhance the quality of extracted lensless images, applied at the final stage of the primary feature extractor. However, this attention mechanism is limited to high-level features and overlooks relational information across various frequency variations. Consequently, its performance is not significantly improved, achieving only 82.81% accuracy at the highest resolution.

Table 5. Performance comparison of the proposed model with previous works at different input resolutions evaluated on the FERPlus dataset.

Model	Accuracy on Resolution					
	256 × 256	224 × 224	48 × 48	20 × 20	10 × 10	5 × 5
SCN	-	-	-	78.09	69.62	52.71
DMUE	-	-	-	74.23	59.57	40.99
RUL	-	-	-	79.16	68.69	58.23
MULR	-	-	-	79.57	71.87	61.39
He et al.	-	-	**84.63**	-	-	-
Cho et al.	-	88.45	-	-	-	-
Ma et al. [42]	-	88.81	-	-	-	-
LenslessFET [43]	82.81		-	-	-	-
Jiang et al. [40]	-	**89.64**	-	-	-	-
FER-MOTION	**87.50**	89.46	84.42	**81.53**	**74.16**	**63.52**

Table 6. Comparison of model efficiency between the proposed facial emotion network with other lightweight architectures.

Model	Image size	Parameters	GFLOPS	Speed in FPS on Jetson Nano		Speed in FPS on CPU, i7-12700		Acc on KDEF
				FER	Int	FER	Int	
MobileNetV1	32 × 32	3,236,039	0.0235	19.88	13.24	267.20	29.69	96.15
MobileNetV2	32 × 32	2,266,951	0.0126	13.70	10.20	221.57	28.90	96.59
MobileNetV3 Small	32 × 32	2,949,663	0.0168	11.30	8.84	226.34	20.32	95.58
MobileNetV3 Large	32 × 32	5,127,839	0.0179	9.87	7.92	185.51	27.35	96.75
ShuffleNetV1	32 × 32	973,567	0.0060	28.26	16.58	287.92	29.92	90.08
ShuffleNetV2	32 × 32	4,025,915	0.0201	20.05	13.42	182.98	27.07	96.14
ResNet18	32 × 32	11,198,919	0.0350	8.30	5.75	189.52	27.67	97.12
GhostNet	32 × 32	3,918,680	0.0113	18.26	12.24	198.64	28.63	96.79
FER-MOTION -10	**10 x10**	**595.549**	**0.0070**	**10.34**	**8.17**	**216.32**	**29.21**	**92.10**
FER-MOTION -32	**32 × 32**	**595.549**	**0.0070**	**9.73**	**7.75**	**191.87**	**27.78**	**97.96**
FER-MOTION -224	**224 × 224**	**595.549**	**0.0070**	**8.18**	**6.90**	**34.88**	**15.78**	**97.81**

This evaluation also illustrates that our model obtained high performance in other resolutions with an accuracy of 81.53%, 74.16%, and 63.52% at 20 × 20, 10 × 10, and 5 × 5, respectively. On the other hand, He et al.'s model slightly outperforms our model at 48 × 48 resolution with an accuracy of 84.63%, making it marginally the best at this resolution. Nonetheless, the proposed model remains competitive at this intermediate level. The FER-MOTION demonstrates superior performance across various resolutions, excelling at higher input image sizes.

4.3 Efficiency evaluation and implementation of real scenario

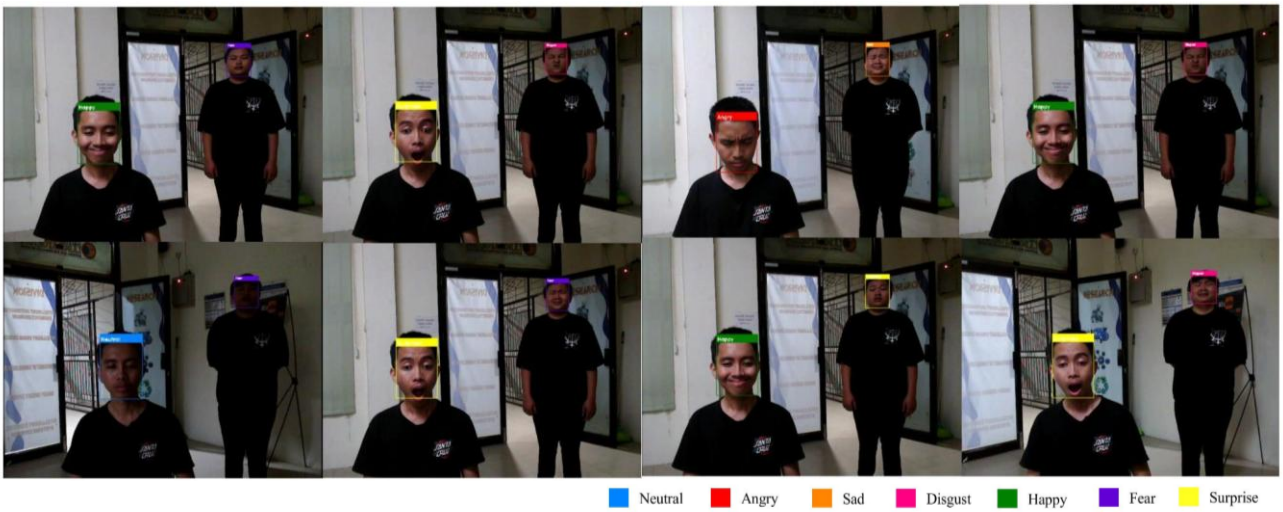
CNN models have demonstrated exceptional performance in recognizing facial expressions due to their ability to learn through weighted spatial filter operations. This approach typically involves a deep architecture with numerous convolutional layers, significantly increasing computational demands. Consequently, the extensive operations can lead to a reduction in data processing speed. In our proposed work, we investigate the efficiency of the developed model by examining the number of parameters, computational complexity, and processing speed and comparing it with existing lightweight architectures. To evaluate the practicality of our model, we conducted speed tests on a Jetson Nano as a commonly used low-cost device in applications such as robotics, IoT, smart home, and industrial systems. These tests are intended to demonstrate the model's reliability in maintaining speed on resource-constrained hardware.

The testing stage demonstrates that the proposed model generates fewer parameters and requires less computational effort than competing models, as detailed in Table 6. All experiments were conducted at the exact input resolution of 32 × 32 for a fair comparison. The FER-MOTION consistently maintained a parameter count of 595 million and 7 GFLOPs (Giga Float Point Operations). Despite its strengths in performance and efficiency, the facial expression (FER) model speed lags behind other lightweight architectures (except ResNet-18), achieving only 9.73 Frame Per Second (FPS) at the 32 × 32 resolution. It is primarily due to the depthwise convolution operations within our network, which necessitate parallel processing memory. Besides, the Keras framework processes these operations sequentially, resulting in slower feature extraction and processing. Nevertheless, this limitation does not significantly impact the model's practical application, particularly when integrated with a face detection system (int). The facial emotion detection system remains reliable, operating at 7.75 FPS with satisfactory performance effectiveness. To evaluate the model's performance on a low-cost device designed for real-time speed, we also deployed the proposed model on a CPU-based system equipped with an Intel Core i7-12700 processor (4.4 GHz) and 16 GB of RAM. Notably, this device is more affordable than the Jetson Nano. The proposed model achieved an impressive speed of 191.87 FPS, outperforming ResNet-18 and ShuffleNetv2. However, the extensive use of depthwise operations and the branching in the convolution layers introduced a bottleneck that slightly hindered the

network's efficiency, a drawback of the proposed algorithm.

The visualization presented in Fig. 5(a) serves to demonstrate the reliability of our model when applied to real-world scenarios in normal illumination. This experiment covers small, medium, and large face scenarios, using the KDEF dataset as the knowledge base. To ensure accurate face recognition, a face detection algorithm [44] was utilized to isolate and focus on face regions, generating face patches for analysis. The proposed system employs models at three distinct resolution levels to optimize performance across various face sizes. For face patches smaller than 30 × 30, models trained at a resolution of 20 × 20 are used. For medium-sized faces ranging from 30 × 30 to 100 × 100, models trained at 32 × 32 are applied. Larger faces exceeding 100 × 100 are processed using models trained at 224 × 224. This multi-resolution approach enables the system to accurately predict faces at different scales, effectively recognizing faces even at a distance of 4 meters from the camera in the case of small faces. In order to assess the reliability and capability of the proposed model, we conducted tests under various lighting conditions, as illustrated in Fig. 5(b). The model successfully recognizes facial expressions in the top row of the image, even under limited lighting in the facial area. However, the bottom row highlights prediction errors in low-illumination scenarios. The scarcity of facial gesture information challenges the model in accurately predicting sad expressions. Furthermore, the model struggles significantly when the face is subjected to extremely low illumination.

(a)



(b)

Figure. 5 Visualization of test results in real-world scenarios across multiple facial input resolutions. The scenarios are conducted under normal illuminance (a) and abnormal illuminance (b)

5. Conclusions and future works

This work proposes FER-MOTION, a facial expression recognition system that identifies emotions across multi-resolution inputs. The deep learning network effectively distinguishes facial texture features while maintaining efficiency, irrespective of variations in input dimensions. A Large Receptive Residual Network (LR2) can discriminate between critical features and trivial information without requiring extensive computational resources. Additionally, the system incorporates two enhancement modules to bolster feature extraction capabilities. The Large Context Enhancement module refines LR2 by enhancing focus on gesture context, improving feature filtering. Simultaneously, the Efficient Pyramid Enhancement module emphasizes essential features by representing diverse receptive areas. The model analysis demonstrates that the proposed modules significantly enhance network performance, minimizing the need for large numbers of trained parameters and computational resources. Comparative evaluations against previous work across different input resolutions show that the proposed recognizer consistently performs better. Although the model operates slower than other lightweight CNN architectures, it utilizes fewer parameters and exhibits reduced computational complexity. This deep learning model represents a significant scientific contribution as a novel network for recognizing human facial emotions across different resolution scales. Furthermore, the proposed modules, such as Large Receptive Residual Network, Large Context Enhancement, and Efficient Pyramid Enhancement, offer valuable recommendations for designing efficient and effective deep learning networks to enhance lightweight feature extraction performance. Future research may focus on optimizing the loss function to improve recognition accuracy by exploring focal loss approaches. Additionally, implementing the system on a robot would provide valuable insights into real-world scenarios, particularly in environments with dynamic illumination and jitter effects.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, M. D. Putro, Wahyono, and D. C. Hernández; methodology, M. D. Putro; software, J. Hariyono, and O. A. Lantang; validation, Wahyono, and M. D. Putro; data processing, O. A. Lantang; writing—original draft preparation, M. D. Putro, J. Hariyono, and O. A. Lantang; writing—review and editing, Wahyono and D. C. Hernández; visualization, M. D. Putro; supervision, Wahyono; project administration, O. A. Lantang.

Acknowledgments

This research was supported by Directorate of Research, Technology, and Community Service Ministry of Education, Culture, Research and Technology, Indonesia in 2024 Fundamental Research scheme under the Grant No. 084/E5/PG.02.00.PL/2024 and 1891/UN12.13/LT/2024.

References

[1] F. Ma, B. Sun and S. Li, “Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion”, *IEEE Transactions on Affective Computing*, Vol. 14, No. 2, pp. 1236-1248, 2023, doi: 10.1109/TAFFC.2021.3122146.

[2] Z. Zhao, Q. Liu and S. Wang, “Learning Deep Global Multi-Scale and Local Attention Features for Facial Expression Recognition in the Wild”, *IEEE Transactions on Image Processing*, Vol. 30, pp. 6544-6556, 2021, doi: 10.1109/TIP.2021.3093397.

[3] M. D. Putro, D. -L. Nguyen and K. -H. Jo, “A Fast CPU Real-Time Facial Expression Detector Using Sequential Attention Network for Human-Robot Interaction”, *IEEE Transactions on Industrial Informatics*, Vol. 18, No. 11, pp. 7665-7674, 2022, doi: 10.1109/TII.2022.3145862.

[4] Y. Yan, Z. Zhang, S. Chen, and H. Wang, “Low-resolution facial expression recognition: A filter learning perspective”, *Signal Processing*, Vol. 169, p. 107370, 2020.

[5] Caifeng Shan, Shaogang Gong and P. W. McOwan, “Recognizing facial expressions at low resolution”, *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, Como, Italy, pp. 330-335, 2005, doi: 10.1109/AVSS.2005.1577290.

[6] Bodavarapu PNR, Srinivas PVVS, “Facial expression recognition for low resolution images using convolutional neural networks and denoising techniques”, *Indian Journal of Science and Technology*. 14(12): 971-983. 2021. https://doi.org/10.17485/IJST/v14i12.14

[7] K. Karilingappa, D. Jayadevappa, and S. Ganganna, “Human emotion detection and classification using modified Viola-Jones and

convolution neural network", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 12, No. 1, pp. 79-86, 2023.

[8] W. Liu, T. Zhang, S. Xu, Q. Chang and Y. Cui, "PanoDetNet: Multi-Resolution Panoramic Object Detection With Adaptive Feature Attention", *IEEE Access*, Vol. 12, pp. 104300-104316, 2024, doi: 10.1109/ACCESS.2024.3435764.

[9] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-MobileNet: An approach of facial expression recognition", *Alexandria Engineering Journal*, Vol. 61, No. 6, pp. 4435-4444, 2022.

[10] B. Hdioud and M. E. H. Tirari, "Facial expression recognition of masked faces using deep learning", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 12, No. 2, pp. 921-930, 2023.

[11] F. Xue, Q. Wang, Z. Tan, Z. Ma and G. Guo, "Vision Transformer With Attentive Pooling for Robust Facial Expression Recognition", in *IEEE Transactions on Affective Computing*, Vol. 14, No. 4, pp. 3244-3256, 2023, doi: 10.1109/TAFFC.2022.3226473.

[12] J. Jien, A. Baharum, S. Wahab, N. Saad, M. Omar, and N. Noor, "Age-based facial recognition using convoluted neural network deep learning algorithm", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 9, No. 3, pp. 424-428, 2020.

[13] H. Benradi, A. Chater, and A. Lasfar, "A hybrid approach for face recognition using a convolutional neural network combined with feature extraction techniques", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 12, No. 2, pp. 627-640, 2023.

[14] R. Jiménez-Moreno and R. A. Castillo, "Deep learning speech recognition for residential assistant robot", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 12, No. 2, pp. 585-592, 2023.

[15] M. D. Putro, D. -L. Nguyen and K. -H. Jo, "An Efficient Face Detector on a CPU Using Dual-Camera Sensors for Intelligent Surveillance Systems", *IEEE Sensors Journal*, Vol. 22, No. 1, pp. 565-574, 2022, doi: 10.1109/JSEN.2021.3128389.

[16] Z. Gong, A. P. French, G. Qiu and X. Chen, "CTranS: A Multi-Resolution Convolution-Transformer Network for Medical Image Segmentation", *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece, 2024, pp. 1-5, doi: 10.1109/ISBI56570.2024.10635192.

[17] Y. Wang, S. Chen, H. Bian, W. Li and Q. Lu, "Deep Multi - Resolution Network for Real-Time Semantic Segmentation in Street Scenes", *2023 International Joint Conference on Neural Networks (IJCNN)*, Gold Coast, Australia, pp. 01-08, 2023, doi: 10.1109/IJCNN54540.2023.10191758.

[18] J. H. Kim, A. Poulose and D. S. Han, "CVGG-19: Customized Visual Geometry Group Deep Learning Architecture for Facial Emotion Recognition", *IEEE Access*, Vol. 12, pp. 41557-41578, 2024, doi: 10.1109/ACCESS.2024.3377235.

[19] S. M. Hassan and A. K. Maji, "Pest Identification Based on Fusion of Self-Attention With ResNet", *IEEE Access*, Vol. 12, pp. 6036-6050, 2024, doi: 10.1109/ACCESS.2024.3351003.

[20] S. Subudhiray, H. K. Palo, and N. Das, "K-nearest neighbor based facial emotion recognition using effective features", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 12, No. 1, pp. 57-65, 2023.

[21] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 8, pp. 2011-2023, 2020, doi: 10.1109/TPAMI.2019.2913372.

[22] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network", *Computer Vision -- ACCV 2022*, 2023, pp. 541-557.

[23] T. Alghamdi and G. Alaghband, "SAFEPA: An Expandable Multi-Pose Facial Expressions Pain Assessment Method", *Applied Sciences*, Vol. 13, No. 12, 2023.

[24] M. Aly, A. Ghallab and I. S. Fathi, "Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model", *IEEE Access*, Vol. 11, pp. 121419-121433, 2023, doi: 10.1109/ACCESS.2023.3325407.

[25] T. Liu, J. Li, J. Wu, B. Du, J. Chang and Y. Liu, "Facial Expression Recognition on the High Aggregation Subgraphs", *IEEE Transactions on Image Processing*, Vol. 32, pp. 3732-3745, 2023, doi: 10.1109/TIP.2023.3290520.

[26] M. Khan, A. E. Saddik, M. Deriche and W. Gueaieb, "STT-Net: Simplified Temporal Transformer for Emotion Recognition", in *IEEE Access*, Vol. 12, pp. 86220-86231, 2024, doi: 10.1109/ACCESS.2024.3413136.

[27] W. Dou, K. Wang and T. Yamauchi, "Face Expression Recognition With Vision

Transformer and Local Mutual Information Maximization", *IEEE Access*, Vol. 12, pp. 169263-169276, 2024, doi: 10.1109/ACCESS.2024.3496506.

[28] J. Aina, O. Akinniyi, M. M. Rahman, V. Odero-Marah and F. Khalifa, "A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition", in *IEEE Access*, Vol. 12, pp. 91410-91425, 2024, doi: 10.1109/ACCESS.2024.3421376.

[29] B. Lee, K. Ko, J. Hong and H. Ko, "Hard Sample-aware Consistency for Low-resolution Facial Expression Recognition", *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp. 198-207, 2024, doi: 10.1109/WACV57701.2024.00027.

[30] M. D. Putro, J. Litouw, and V. C. Poekoel, "Low-resolution facial emotion recognition on low-cost devices", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 13, No. 2, pp. 2201-2211, 2024.

[31] L. Lo, B. -K. Ruan, H. -H. Shuai and W. -H. Cheng, "Modeling Uncertainty for Low-Resolution Facial Expression Recognition", *IEEE Transactions on Affective Computing*, Vol. 15, No. 1, pp. 198-209, 2024, doi: 10.1109/TAFFC.2023.3264719.

[32] M. D. Putro, D. -L. Nguyen and K. -H. Jo, "A Fast CPU Real-Time Facial Expression Detector Using Sequential Attention Network for Human-Robot Interaction", *IEEE Transactions on Industrial Informatics*, Vol. 18, No. 11, pp. 7665-7674, 2022, doi: 10.1109/TII.2022.3145862.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 4510-4520, 2018, doi: 10.1109/CVPR.2018.00474.

[34] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition", *IEEE Transactions on Image Processing*, Vol. 28, No. 1, pp. 356-370, 2019, doi: 10.1109/TIP.2018.2868382.

[35] M. G. Calvo and D. Lundqvist, "Facial expressions of emotion (KDEF): Identification under different display-duration conditions", *Behavior Research Methods* , Vol. 40, No. 1, pp. 109-115 2008, doi: 10.3758/BRM.40.1.109.

[36] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution", In: *Proc. of the 18th ACM International Conference on Multimodal Interaction*, pp. 279-283, 2016, doi: 10.1145/2993148.2993165.

[37] M. Aly, A. Ghallab and I. S. Fathi, "Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model", *IEEE Access*, Vol. 11, pp. 121419-121433, 2023, doi: 10.1109/ACCESS.2023.3325407.

[38] S. Cho and J. Lee, "Learning Local Attention With Guidance Map for Pose Robust Facial Expression Recognition", *EEE Access*, Vol. 10, pp. 85929-85940, 2022, doi: 10.1109/ACCESS.2022.3198658.

[39] M. Sun et al., "Attention-Rectified and Texture-Enhanced Cross-Attention Transformer Feature Fusion Network for Facial Expression Recognition", in IEEE Transactions on Industrial Informatics, Vol. 19, No. 12, pp. 11823-11832, 2023, doi: 10.1109/TII.2023.3253188.

[40] C. -S. Jiang, Z. -T. Liu, M. Wu, J. She and W. -H. Cao, "Efficient Facial Expression Recognition With Representation Reinforcement Network and Transfer Self-Training for Human-Machine Interaction", *IEEE Transactions on Industrial Informatics*, Vol. 19, No. 9, pp. 9943-9952, 2023, doi: 10.1109/TII.2022.3233650.

[41] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets", *IEEE Access*, Vol. 12, pp. 45543-45559, 2024, doi: 10.1109/ACCESS.2024.3380847.

[42] F. Ma, B. Sun and S. Li, "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion", *IEEE Transactions on Affective Computing*, Vol. 14, No. 2, pp. 1236-1248, 2023, doi: 10.1109/TAFFC.2021.3122146.

[43] J. Yang, M. Zhang, X. Yin, K. Li and H. Yue, "Lensless Sensing of Facial Expression by Transforming Spectral Attention Features", *IEEE Transactions on Instrumentation and Measurement*, Vol. 73, pp. 1-13, Art No. 5014213, 2024, doi: 10.1109/TIM.2024.3375987.

[44] M. D. Putro, A. Priadana, D. -L. Nguyen and K. -H. Jo, "A Faster Real-time Face Detector Support Smart Digital Advertising on Low-cost Computing Device", In: *Proc. of 2022*

IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Sapporo, Japan, pp. 171-178, 2022, doi: 10.1109/AIM52237.2022.9863289.